我国战略生物资源大数据及应用

梁 卓[1] 褚 鑫[2] 曾 艳[1] 周 桔[1] 马俊才[3*]

1 中国科学院科技促进发展局 北京 100864
2 中国科学院重大科技任务局 北京 100864
3 中国科学院微生物研究所 北京 100101

摘要 生物资源是国家重要的战略性资源。组学时代的到来催生数据量级的增长将推动生命科学研究向“数据密集型科学”的新范式转变。因此，生物资源的大数据与生物资源本身一样，也已成为国家战略资源，成为国际科技与产业竞争热点和战略制高点。目前，我国亟待加快建设国家级战略生物资源大数据共享平台，建立全面支撑生物资源保存、利用、开发、挖掘和可持续利用的技术体系，形成包括数据质控流程、数据存储中心、数据共享中心和数据转化中心的可对海量数据进行有效管理、高效分析和可用、易用的综合数据系统，支撑我国生物资源前沿领域科学发现和产业创新发展。中国科学院战略生物资源信息平台配合中国科学院战略生物资源服务网络的建设，形成了包括数据管理规范、数据汇集平台、数据门户和数据可视化系统在内的一套完整的数据生态，从而有效地促进了中国科学院战略生物资源数据的集成、共享、挖掘和利用。目前，该平台已经覆盖中国科学院植物园、标本馆、生物遗传资源、实验动物、生物多样性监测网络 5 类资源的 42 家单位，集成 6 981 465 条数据，且全部数据资源都可以通过数据门户进行共享。

关键词 生物资源，信息化，大数据

DOI 10.16418/j.issn.1000-3045.2019.12.009

1 生物资源是国家的战略性资源

生物资源是指对人类具有现实和潜在价值的基因和物种的总和，包括植物、动物、微生物资源和人类遗传资源。生物资源是人类繁衍和发展最基本的物质基础，也是地球上最宝贵的财富[1]。生物资源是生物产业、现代农业和生命科学研究的源头与基础。特色植物、资源植物可能为整个国家的农业带来转型。丰富多样的微生物为人类解决能源、环境危机提供了重

*通讯作者

资助项目：中国科学院战略生物资源专项（KFJ-BRP-006），中国科学院战略性先导科技专项（XDA19050301），国家重点研发计划（2016YFB1000605、2016YFC1200801、2016YFC0901702、2017YFC1201202），国家科技基础条件平台专项课题（2018DDJ1ZZ08）

修改稿收到日期：2019年11月20日

要平台[2]。干细胞基于其自我更新和分化潜能，为人类许多重大疾病的根治带来希望。模式动物与实验动物赋予了人类揭示生命本质的启示和手段，而各类工具酶的应用则是现代分子遗传学发展和基因工程的基础。生物资源是国家的战略性资源已成为各行各业的共识[1]。

2 战略生物资源大数据是未来国际科技与产业竞争热点和战略制高点

新技术带来的数据激增以及数据管理要求的提高，推动着生命科学向大数据驱动发现的维度发展[3]。21 世纪初“人类基因组计划”的完成标志着生命科学已经进入大数据时代，一种新的生命科学研究范式出现[4]。大数据革命将改变我们的生活和商业模式[5]。

科学数据是科技创新和社会经济发展的重要基础。大数据的应用将成为未来竞争的基础。标本馆的收藏及其保存的数据不仅对于传统的分类学和系统学研究很有价值，而且对于生态学、生物工程、食品安全以及科学收藏的人类社会和文化元素也很有价值[6]。植物学数据可用作开发统计模型的训练数据，以预测变化对生物体的影响。这些模型可用作保护和政策工具，以减轻全球环境变化对生物多样性和粮食安全的影响。基于原始生物多样性数据建立的物种分布模型也可用于更好地发现和预测生物入侵。整合新生物和古生物多样性数据，与在资源中发现的基于文献的发生数据相联系，可以帮助回答当今全球变化研究的深层问题。科学数据作为国家科技创新发展和经济社会发展的重要基础性战略资源，已经成为全社会的高度共识。与生物资源本身一样，生物资源信息数据也已成为国家重要战略资源，成为国际科技与产业竞争热点和战略制高点。大数据基础设施提供大规模数据收集、存储以及处理和分析的能力，是国家核心竞争力的重要组成部分。

因此，建设一个集中的国家级战略生物资源大数据基础设施成为当务之急。这种基础设施应实现卓越的安全数据存储能力、标准化数据处理和质量控制、跨多种类型的系统数据集成，以及深入的数据挖掘和有效的数据共享。利用这种基础设施作为发射台，我们将能够不断提高国家安全数据存储、信息共享、技术创新、标准化系统改进、知识产权增值和生物数据高效利用的能力。

3 国内外战略生物资源数字化建设的现状与进展

3.1 发达国家高度重视对生物资源的信息化建设和管理

*在过去几十年中，发达国家创建、存储和连接生物多样性数据库的举措激增。*例如，英国国家生物多样性网络门户①是一个收集、分类、分析和传播英国生物多样性数据的组织。欧洲建立了生物多样性数据中心②，爱尔兰和比利时也倡导了目的类似的国家计划③[7]。长期监测项目，如美国的长期生态研究网络项目（LTER）④、美国国家科学基金会（NSF）出资 4 亿美元建立的国家生态观测网（NEON）⑤以及澳大利亚的陆地生态系统研究网络（TERN）⑥等，正在提供长期标准化的生态和生物数据集的开放访问[8]。更雄心勃勃的国际基础设施寻求将各国和各大洲的数

① http://data.nbn.org.uk/.
② http://www.eea.europa.eu/themes/biodiversity/dc.
③ http://www.biodiversityireland.ie/ and http://data.biodiversity.be/.
④ http://lternet.edu/.
⑤ http://www.neonscience.org/.
⑥ http://www.tern.org.au/.

据库连接起来。全球生物多样性信息设施（GBIF）⑦是一个跨国、开放的数据基础设施，用于汇集当地和不同来源的生物多样性信息，目前已包括10亿多条物种分布信息。该设施“允许任何地方的任何人通过互联网访问跨越国界共享的关于地球上所有类型生命的数据”。同样，由美国NSF资助的生态信息技术平台DataONE（地球数据观测网络）⑧备受瞩目；该项目目标是广泛收集、存储地球和环境有关数据，在普及应用基础上，创造新的知识[9]。

馆藏数字化已成为世界自然历史馆藏的关键活动之一。通过NSF资助推进的生物多样性数字化（ADBC）项目，美国的保藏物数字化得到了加强。到2020年，美国博物馆中数字化的标本总数可能超过1亿件[10]。

近年来，生物多样性信息学快速发展，全球和区域水平的生物多样性数据库不断建立和完善[11]。若干国家水平的数据库，如澳大利亚生物多样性信息系统（ALA）⑨已有超过4 606万条记录，美国标本数字化平台（iDigBio）⑩有超过9 500万条的数字化标本共享。全球生物信息的主要数据由美国的国家生物技术信息中心（NCBI）和欧洲生物信息研究所（EBI）等建立的数据库控制[12]。

2014年启动的世界植物在线（WFO）项目扩大了数字化工作和虚拟植物标本馆的范围。WFO是根据《全球植物保护战略》2020年的主要目标开发的一个与世界其他30个机构合作的项目，旨在制作并呈现世界在线植物区系。这些数据通过虚拟植物标本馆呈现，并提供给WFO门户网站；所有参与机构将向该网站提供其拥有的数字化植物描述。

2015年，中国作为代表提出了亚洲生物多样性数字化计划（Mapping Asia Plants，MAP）。MAP旨在建立亚洲植物学信息的大数据在线平台，为亚洲植物多样性保护与研究提供综合性基础信息和跨学科数据挖掘环境[13]。与MAP类似的项目为数不多，全球尺度上有Map of Life（MOL）⑪。MOL旨在搜集和整合全球物种分布及其动态变化的数据和知识，为生物多样性教育、保护、研究和科学决策服务。其理念是把生物多样性画在图上（putting biodiversity on the map）。在区域尺度上，植物学信息与生态网络（Botanical Information and Ecology Network，BIEN）⑫提供了很好的参考案例。

3.2 我国积极推动科技资源的整合共享与高效利用

为促进科技资源的有效利用和共享，我国科学技术部启动了国家科技基础条件平台建设工作。自然科技资源和科学数据共享服务属于平台建设的主要内容，植物种质资源、动物种质资源、微生物菌种资源、人类遗传资源、实验生物材料、生物标本、岩石矿物和化石标本等八大类自然资源的整合集成与共享成为自然科技资源平台建设的重点。科学数据资源建设重点整合、集成各部门、各地方、各单位的科学数据资源，充分利用国际科学数据资源，抢救离散科学数据资源，开发系列数据集和产品，构建面向全社会的网络化、智能化的科学数据管理与共享服务体系。

2018年2月，为深入实施创新驱动发展战略，规范管理国家科技资源共享服务平台，推进科技资源向社会开放共享，科学技术部、财政部印发《国家科技资源共享服务平台管理办法》。2018年4月，国务院办公厅印发《科学数据管理办法》，对战略

⑦ http://www.gbif.org.
⑧ http://www.dataone.org.
⑨ http://www.ala.org.au.
⑩ http://www.idigbio.org.
⑪ http://www.mol.org.
⑫ http://bien.nceas.ucsb.edu/bien.

生物资源数据的规范管理具有重要的指导意义。2019年6月5日，为落实《科学数据管理办法》和《国家科技资源共享服务平台管理办法》的要求，规范管理国家科技资源共享服务平台，科学技术部、财政部对原有国家科技资源共享服务平台开展了优化调整工作。通过部门推荐和专家咨询，形成“国家微生物数据中心”等20个国家科学数据中心以及“国家重要野生植物种质资源库”等30个国家生物种质与实验材料资源库。

4 中国科学院在资源整合共享和数据管理利用方面的进展

中国科学院作为国家战略科技力量，历来十分重视生物资源的保存、利用与共享。在中国科学院战略生物资源服务计划整体框架下，依托中国科学院微生物研究所成立了中国科学院战略生物资源信息中心。该中心以中国科学院在生物资源领域的雄厚积淀为基础，全面整合中国植物园联盟、国家标本馆体系、全国微生物和细胞资源应用网络、模式与特色动物实验平台利用联盟、中国生物多样性监测与研究网络资源数据信息，旨在实现数据共享，开展数据挖掘，提升数据价值，支撑战略生物资源领域科学研究以及社会应用。同时，该平台通过开展战略生物资源领域国内外信息搜集整理和分析，提供政策咨询参考，支撑科学决策，将保存的生物遗传资源转化为数据资源，实现实物资源信息的高效共享服务，并为国家相关部门和国际履约提供技术支撑。

为规范和加强中国科学院战略生物资源服务计划的项目数据汇交工作，促进项目数据的共享，中国科学院科技促进发展局牵头制定了《中国科学院战略生物资源服务计划数据管理及标准规范》，其内容包括“数据管理及汇交办法”和“标准规范”两部分。“数据管理及汇交办法”对服务计划中的数据汇交计划、数据汇交内容、数据汇交流程、汇交数据管理、数据使用与知识产权等内容进行了明确的规定。“标准规范”部分按照不同资源类型及其子类，对需要汇交数据的详细内容、数据格式和样例给出了详细的要求。为促进生物资源数据被尽可能广泛和自由地共享和使用，实现数据的规范管理、有效利用和增值服务提供了参考和依据。

中国科学院战略生物资源信息中心开发了我国战略生物资源对外信息共享门户和综合服务平台[13]。作为我国重要战略生物资源的基础数据平台，该平台提供统一的检索入口和并行综合查询，发挥其核心数据目录的作用，为我国的生命科学研究以及生物技术的发展提供重要的数据支撑。依据中国科学院战略生物资源服务计划数据标准规范建立的在线目录资源数据管理平台，可以按照各个资源领域及其数据提交单位设立不同的管理权限，并可通过基于网页方式的数据管理平台，建立多级的数据上传、同步和发布等功能，从而有效支撑了全院各资源领域、各资源中心的数据汇交、实时发布和在线统计。利用数据管理工具，形成了各保藏单位、工委会、总中心的三级数据汇交体系，并汇总形成中国科学院战略生物资源目录（图1）。利用该目录，可以数据按照来源、物种、生存状态、共享方式等多种类型进行统计，并能进行多样化的在线检索和数据展示。

中国科学院战略生物资源服务网络信息平台致力于涵盖中国科学院内标本馆、植物园、典型培养物、实验动物和生物多样性监测网络，进行整体布局和顶层设计（表1）。该平台建立了一系列生物资源数据的收集、保护、研究、管理与利用等基础数据的采集、加工、处理与交换的标准规范；整合了中国科学院生物资源数据库，彻底理清了中国科学院生物资源

⑬ http://www.casbrc.org.

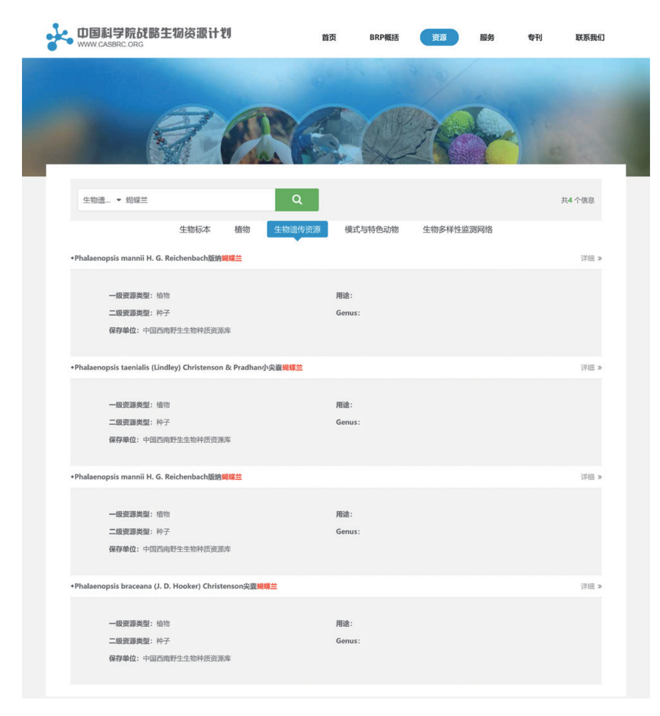


图 1　中国科学院战略生物资源在线目录

表 1　中国科学院战略生物资源数据发布情况

资源库	数据总量	馆藏单位数量
生物标本	6 425 949	18
植物园	144 980	13
生物遗传	25 538	10
模式与特色动物	1 173	19
观测网	88 568	4
总计	6 686 208	64

的家底；建立了统一的数据服务平台和完备的数据交换与共享机制，建立了全院统一的资源目录和信息门户，包括各单位馆藏的元数据信息、全院综合资源目录及信息门户，从而提高中国科学院生物资源保护、研究与管理水平，促进战略生物资源面向社会的应用和服务水平。

5 思考及讨论

在大数据的背景下，未来的生物资源研究必将朝着形成一个全方位的生物资源研究、开发与应用的网络的方向发展，各个环节的联系更加紧密，对数据应用也必将提出更高的要求。系统、全面、科学地搜集、整理我国生物资源信息，不仅可以将其用于生物多样性保护和生物资源的合理开发和利用，而且可以应用到以生物学为基础的各个方面，推动中国生命科学研究的快速发展，进而影响生物经济的各个领域。因此，建立战略生物资源综合信息服务平台不仅是基础科学发展的客观要求和必然趋势，也是国民经济和社会可持续发展的需要。随着我国国民经济建设的飞速发展，各行各业要求生物学家提供越来越多的生物资源信息。中国科学院战略生物资源综合信息服务平台通过整体布局和顶层设计，为生物资源的系统化、集约化保存、研究和功能评价提供全方位的信息支撑；通过建立信息集成和服务平台，向全社会开放、共享资源，凸显生物资源对经济社会发展的重要支撑服务作用。未来，中国科学院还将以该信息平台为基础，开展生物资源搜集和功能研究的广泛的国际合作。通过开展联合测序、数据分析等共同研究，对生物资源进行研究及功能评价，拓宽我国资源搜集渠道，为生物技术应用提供基础。

参考文献

1 段子渊, 黄宏文, 刘杰, 等. 保存国家战略生物资源的科学思考与举措. 中国科学院院刊, 2007, 22(4): 284-291.

2 刘柳, 马俊才. 国际微生物大数据平台的应用与启示. 中国科学院院刊, 2018, 33(8): 846-852.

3 Thessen A, Patterson D. Data issues in the life sciences. ZooKeys, 2011, 150: 15-51.

4 Hood L, Rowen L. The human genome project: Big science transforms biology and medicine. Genome Medicine, 2013, 5(9): 79.

5 Wang Z, Wang Z, Li Y. Strategic planning for national biomedical big data infrastructure in China. Quantitative Biology, 2017, 5(3): 272-275.

6 James S A, Soltis P S, Belbin L, et al. Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in Plant Sciences, 2018, 6(2): e1024.

7 Devictor V, Bensaude-Vincent B. From ecological records to big data: The invention of global biodiversity. History and Philosophy of the Life Sciences, 2016, 38: 13.

8 Ladeau S L, Han B A, Rosi-Marshall E J, et al. The next decade of Big Data in ecosystem science. Ecosystems, 2017, 20(2): 274-283.

9 Michener W K, Allard S, Budden A, et al. Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics, 2012, 11: 5-15.

10 Thiers B M, Tulig M C, Watson K A. Digitization of the New York Botanical Garden Herbarium. Brittonia, 2016, 68(3): 324-333.

11 Wu L, Sun Q, Desmeth P, et al. World data centre for microorganisms: An information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Research, 2017, 45(D1): D611-D618.

12 马克平, 朱敏, 纪力强, 等. 中国生物多样性大数据平台建设. 中国科学院院刊, 2018, 33(8): 838-845.

13 马克平. 亚洲植物多样性数字化计划. 生物多样性, 2017, 25(1): 1-2

Big Data and Application of Strategic Biological Resources in China

LIANG Zhuo[1] CHU Xin[2] ZENG Yan[1] ZHOU Ju[1] MA Juncai[3*]

（1 Bureau of Science and Technology for Development, Chinese Academy of Sciences, Beijing 100864, China;

2 Bureau of Major R&D Programs, Chinese Academy of Sciences, Beijing 100864, China;

3 Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China）

Abstract It is well understood among industries that biological resources are the strategic resources of the nation. The increase in the magnitude of the data generated by the arrival of the omniscience era will promote the transformation of life science research into a new paradigm of “data-intensive science”. Therefore, the big data of biological resources has become a national strategic resource as well as the focus of international technology competition and strategic commanding heights. At present, to establish a national infrastructure for biological resources is urgently needed. This infrastructure incorporates a technical system that supports the conservation, utilization, development, mining and sustainable use of biological resources and an integrated data system which can effectively manage, analyze and use massive data. These services will contribute greatly to support the scientific discovery and industrial innovation development in the frontier of biological resources of our nation. Biological resources information platform has cooperated with Biological Resources Programme, Chinese Academy of Sciences (CAS-BRP) to form a complete data system including data management specifications, data collection platform, data portal and data visualization system, which effectively motivate the integration, sharing, mining and utilization of strategic biological resources data of CAS. The data system covers 6 981 465 data of material resources including botanical gardens, biological specimen, biological genetic resources, laboratory animals resouces and biodiversity observation and research network from 42 research institutes of the Chinese Academy of Sciences. All resources can be shared through the data portal www.casbrc.

Keywords biological resources, visualization, big data

* Corresponding author



梁 卓 中国科学院科技促进发展局生物技术处业务主管，主要从事生物技术领域科研管理及成果转化等工作。E-mail: liangzhuo@cashq.ac.cn

LIANG Zhuo Program Officer of Bureau of Science and Technology for Development, Chinese Academy of Sciences (CAS). He is mainly responsible for biotechnology projects management, tech-transfer etc. E-mail: liangzhuo@cashq.ac.cn



马俊才 中国科学院微生物研究所微生物资源与大数据中心主任，世界微生物数据中心主任，世界微生物菌种保藏联合会（WFCC）理事会执委，中华人民共和国科学技术部人类遗传资源管理专家委员会委员。国家“863”计划“微生物数字化信息系统集成关键技术的研发”项目首席科学家。主要研究领域包括：微生物资源和生物技术领域信息化、基于云环境的微生物大数据管理和分析平台。E-mail: ma@im.ac.cn

MA Juncai Received Ph.D. degree from Department of Bio-resources of Mie University in Japan, now he is the director of The Center for Microbial Resource and Big Data, Institute of Microbiology, Chinese Academy of Sciences (CAS). He is also the director of WFCC-MIRCEN World Data Center of Microorganisms (WDCM), and Board Member of World Federation of Culture Collections (WFCC). He is the Principle Investigator of National High Technology Research and Development Program “Key Technology Researches on Microbial Digital Resources Information System”. His research fields cover informalization of microbial and biotechnology resource, cloud based big data management, and analysis system of microbial resources. E-mail: ma@im.ac.cn

■责任编辑：岳凌生

参考文献（双语版）

1 段子渊, 黄宏文, 刘杰, 等. 保存国家战略生物资源的科学思考与举措. 中国科学院院刊, 2007, 22(4): 284-291.
Duan Z Y, Huang H W, Liu J, et al. Considerations and measures on conservation of national strategic bio-resources. Bulletin of Chinese Academy of Sciences, 2007, 22(4): 284-291. (in Chinese)

2 刘柳, 马俊才. 国际微生物大数据平台的应用与启示. 中国科学院院刊, 2018, 33(8): 846-852.
Liu L, Ma J C. Application and enlightenment of international big data platform for microorganism. Bulletin of Chinese Academy of Sciences, 2018, 33(8): 846-852. (in Chinese)

3 Thessen A E, Patterson D J. Data issues in the life sciences. ZooKeys, 2011(150): 15-51.

4 Hood L, Rowen L. The human genome project: Big science transforms biology and medicine. Genome Medicine, 2013, 5(9): 79.

5 Wang Z, Wang Z F, Li Y X. Strategic planning for national biomedical big data infrastructure in China. Quantitative Biology, 2017, 5(3): 272-275.

6 James S A, Soltis P S, Belbin L, et al. Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in Plant Sciences, 2018, 6(2): e1024.

7 Devictor V, Bensaude-Vincent B. From ecological records to big data: The invention of global biodiversity. History and Philosophy of the Life Sciences, 2016, 38(4): 1-23.

8 Ladeau S L, Han B A, Rosi-Marshall E J, et al. The next decade of Big Data in ecosystem science. Ecosystems, 2017, 20(2): 274-283.

9 Michener W K, Allard S, Budden A, et al. Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics, 2012, 11: 5-15.

10 Thiers B M, Tulig M C, Watson K A. Digitization of the New York botanical garden herbarium. Brittonia, 2016, 68(3): 324-333.

11 Wu L, Sun Q, Desmeth P, et al. World data centre for microorganisms: An information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Research, 2016, 45(D1): D611-D618.

12 马克平, 朱敏, 纪力强, 等. 中国生物多样性大数据平台建设. 中国科学院院刊, 2018, 33(8): 838-845.
Ma K P, Zhu M, Ji L Q, et al. Establishing China infrastructure for big biodiversity data. Bulletin of Chinese Academy of Sciences, 2018, 33(8): 838-845. (in Chinese)

13 马克平. 亚洲植物多样性数字化计划. 生物多样性, 2017, 25(1): 1-2.
Ma K P. Mapping Asia plants: A cyberinfrastructure for plant diversity in Asia. Biodiversity Science, 2017, 25(1): 1-2. (in Chinese)